

TMPad: an integrated structural database for helix-packing folds in transmembrane proteins

Allan Lo¹, Cheng-Wei Cheng¹, Yi-Yuan Chiu^{1,2}, Ting-Yi Sung^{1,*} and Wen-Lian Hsu^{1,*}

¹Bioinformatics Laboratory, Institute of Information Science, Academia Sinica, Taipei 115 and ²Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsinchu 300, Taiwan, Republic of China

Received August 15, 2010; Revised November 19, 2010; Accepted November 20, 2010

ABSTRACT

α-Helical transmembrane (TM) proteins play an important role in many critical and diverse biological processes, and specific associations between TM helices are important determinants for membrane protein folding, dynamics and function. In order to gain insights into the above phenomena, it is necessary to investigate different types of helix-packing modes and interactions. However, such information is difficult to obtain because of the experimental impediment and a lack of a well-annotated source of helix-packing folds in TM proteins. We have developed the TMPad (TransMembrane Protein Helix-Packing Database) which addresses the above issues by integrating experimentally observed helix-helix interactions and related structural information of membrane proteins. Specifically, the TMPad offers pre-calculated geometric descriptors at the helix-packing interface including residue backbone/side-chain contacts, interhelical distances and crossing angles, helical translational shifts and rotational angles. The TMPad also includes the corresponding sequence, topology, lipid accessibility, ligand-binding information and supports structural classification, schematic diagrams and visualization of the above structural features of TM helix-packing. Through detailed annotations and visualizations of helix-packing, this online resource can serve as an information gateway for deciphering the relationship between helix-helix interactions and higher levels of organization in TM protein structure and function. The website of the TMPad is freely accessible to the public at <http://bio-cluster.iis.sinica.edu.tw/TMPad>.

INTRODUCTION

α-Helical transmembrane (TM) proteins are a major class of proteins pivotal for many critical biological processes, including signal transductions, bioenergetics, ion transport, cell adhesion and cell-cell recognition (1). It has also been estimated that ~20–30% of a typical genome encodes for proteins with a TM domain (2,3). Despite their biological importance and abundance, the mechanisms by which TM proteins fold into their native states remain elusive due to the limited number of solved structures (4). Therefore, continuous development of data collection and analytical methods that contribute to bridging the sequence-to-structure disparity in TM proteins is highly demanded.

The fold of an α-helical TM protein can be dissected into sets of interacting TM helices, connecting loops and extramembranous domains. In particular, helix-helix interactions remain an important determinant of folding and stabilization by the commonly accepted two-stage model (5,6). Such an interaction is mediated by residue contacts at the helix-packing interfaces and also influenced by protein–lipid and protein–ligand interactions. In particular, how helix-packing contributes to membrane protein assembly has been the subject of many previous works aiming at delineating different helix-packing geometries, sequence motifs and preference of residue contacts. Historically, canonical models of helix-packing such as ‘ridge-into-groove’ (7) and ‘knob-into-hole’ (8) have been proposed for soluble proteins. Later, a comparison between helix-packing in soluble and membrane proteins also reported the remarkable differences in crossing angles, orientation, and packing density (9). In addition, several groups focused on the occurrence of sequence motifs in helix–helix interactions, from the more specific Gly and Ser zippers (10,11), to the degenerate ‘Gly-Ala-Ser’ and ‘Ala-coil’ motifs described by Walters and DeGrado (12). The above over-represented motifs provide clues to an interesting observation that

*To whom correspondence should be addressed. Tel: 886 2 27883799, ext.1711; Fax: 886 2 27824814; Email: tsung@iis.sinica.edu.tw
Correspondence may also be addressed to Wen-Lian Hsu. Tel: 886 2 27883799, ext.1804; Fax: 886 2 27824814; Email: hsu@iis.sinica.edu.tw
Present address:

Allan Lo, Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, Vancouver V5Z 4S6, Canada.

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

small (Ala and Gly) and polar (Ser) residues likely contribute to stability by van der Waals (VDW) and hydrogen bond interactions, respectively. Furthermore, it was also found that non-hydrophobic residues are often buried in the TM core at conserved positions (13). Recently, the above-mentioned helix-packing features were recapitulated and tested by designing a novel peptide sequence which yields high binding specificity to a target TM helix (14). This work demonstrates that the principles of helix-packing in TM proteins can be adapted for gaining insights into membrane protein folding and designing new modulators of protein–protein interactions.

With the above reported progress in this field, one can clearly get a picture that helix-packing in membrane proteins is not governed by a single factor alone; and understanding this phenomenon may require a multi-faceted approach, in which the integration of helix-packing geometries, sequence motifs, contact residues/side-chains, as well as protein–lipid and protein–ligand interactions must be considered. Several databases of related to the above features in membrane protein have been published to-date. One of them is the PDBTM (15), a collection of TM structures from the PDB (16) with computed helical boundaries using the TMDET algorithm (17). Another related work is the OPM database (18) which also contains experimentally determined TM proteins with calculated membrane depths and the orientations of TM proteins in the membranes. In parallel to the above works, the TOPDB database (19) contains topologies of membrane proteins from combined experimental and predicted information. In contrast, two recent methods namely the HIT (20) and the MPlot (21), focused on deriving helix–helix interactions from membrane protein structures in the form of web servers. Both methods allow calculations and visualization of helix–helix interaction features such as interhelical contacts and crossing angles. Another recent work, the MeMotif database (22), provides a collection of linear structural and functional sequence motifs in membrane proteins. Although the above works provide insights into helix–helix interactions in membrane proteins, many other valuable details of helix-packing descriptors are still missing and their relationships to topology, lipid accessibility and ligand information remain unclear. Furthermore, there is still a lack of a comprehensive, well-annotated and integrated repository for existing helix-packing folds in TM proteins.

In the light of addressing the above issues, we have developed the TMPad, an integrated repository of helix-packing folds in α -helical TM proteins. The helix-packing folds in α -helical membrane proteins can be viewed as ensembles of tightly packed helical substructures mediated by helix–helix interactions. The TMPad aims to provide a comprehensive coverage of observed helix–helix interactions with the corresponding sequence, topology, lipid accessibility, ligand and binding site information extracted from experimentally determined membrane protein structures. Furthermore, the TMPad contains annotations of pre-calculated geometric descriptors of each helix-packing interface including residue

backbone/side chain contacts, interhelical distances and crossing angles, helical translational shifts and rotational angles. The TMPad also provides structural classification, schematic diagrams and 3D visualization of the above structural features of TM helix packing. To the best of our knowledge, the TMPad is the first database to provide detailed and integrated annotations of helix-packing structural features in α -helical TM proteins. Given this unique characteristic of the TMPad database, it can be used for deciphering the relationship between helix–helix interactions and higher levels of organization in TM protein structure and function.

DATABASE CONTENT

Data preparation and annotations

The TMPad contains derived data for a comprehensive set of experimentally determined α -helical TM protein structures. In this section, we describe the data collection and preparation process. First, all available TM structures were obtained from the PDBTM and OPM databases. As a pre-processing step, we filtered out the following structures: (i) redundant structures of the same PDB identifiers; (ii) theoretical models and (iii) those without any α -helical TM domains. For each entry in the TMPad, different levels of information are collected in categories including ‘Overview’, ‘Topology’, ‘Helix–helix Interactions’, ‘Lipid Accessibility’ and ‘Ligands’. The details of data preparation in each category are described below:

- (i) Overview: this category presents an overview at the whole protein level including title, experimental details from the PDB annotations, functional and structural classification from the Gene Ontology (GO) (23) and Structural Classification of Proteins (SCOP) databases (24) (if available) and cross references to the PDB, PDBsum (25), PDBTM, JenaLib (26) and UniProt (27) databases, among others. Since there may be chains without a TM domain in a protein, we also list the complete and individual TM-domain chains for comparison. At this level, a quick summary about the protein can be retrieved. The other categories in the TMPad are centered on a selected chain containing at least one TM domain and they are described in details below.
- (ii) Topology: the topology of the protein chain is derived from different sources including the PDBTM, OPM, TOPDB or integrated by consensus predictions. Specifically, we defined the boundaries for any TM helices, reentrant regions, segments of unknown structures based on the annotations in the PDBTM in each protein chain. Furthermore, each TM helix is also classified into one of the following geometric types: ‘linear’, ‘curved’, ‘kinked’ or ‘none of the above’ by using the HELANAL program (28). We also computed the tilt angle (τ) of each TM helix with respect to the membrane normal.

Since the orientation/topology of each chain is often inconsistent or incomplete by the annotations in the above sources, we derived the topology information from OPM, TOPDB or topology predictions based on different criteria. When the topologies were derived from the annotations in OPM/TOPDB with 100 or 70% sequence identities to the target protein chain, we denoted them as ‘By OPM/TOPDB’ or ‘By similarity to OPM/TOPDB’, respectively. In the case when the topology could not be derived from either database, this information was obtained by a majority consensus among the following three prediction servers including SVMtop (29), TOPCONS (30) and Memsat3 (31).

- (iii) Helix–helix interactions: a helix–helix interaction is defined on the basis of the helical backbone and the VDW contacts between a pair of TM helices. Specifically, an interacting helical pair is defined as supporting one or more contacts satisfying the following two distance-based criteria: (i) the distances between C_β-atoms of two residues, one from each helix, are <6 Å and (ii) the atomic distances between any two heavy atoms, one from each helix, are less than the sum of their VDW radii plus 0.6 Å. These identical distance thresholds were also used for defining helix–helix interactions in search of sequence motifs (12). In addition to contacts, we also calculated the interhelical distances by the closest approaches between the helical axes, as well as between the backbone C_α and the heavy atoms of each helical pair. Other important descriptors of helix–helix interactions including crossing angles (Ω), helical translational shifts (per helix: s_1 and s_2) and helical rotational angles (per helix: α and β) are calculated as described below. First, we followed the same definitions by Chothia *et al.* (7) and Lee and Im (32) for defining the principal helical axes. Next, we computed the above geometric constraints (Ω , s_1 , s_2 , α and β) of each helix–helix interaction based on the definitions by Pappu *et al.* (33) which are also shown in Supplementary Figure S1.
- (iv) Lipid accessibility: the lipid accessibility of each protein chain containing at least one helix–helix interaction is calculated using the following procedure. First, we calculated the lipid accessible area (LAA) and the relative lipid accessibility (RLA) of each TM residue in each chain using the NACCESS program (34,35) with the probe radius set to 2.0 Å to mimic the –CH₂ hydrocarbon tail of the lipid bilayer. The RLA describes the relative accessibility of each amino acid through normalization by its reference value. Thus, the RLA of each residue is a real value between 0 and 1.
- (v) Ligands: the ligands category including any ions, natural or synthetic compounds, protein scaffolds, nucleic acids or peptides and the list of their binding residues were extracted from the HETATM fields in the PDB annotations or from the JenaLib database if this information was not available. We also

derived the IUPAC names of each ligand from the JenaLib database and their corresponding common names (if available) from the DrugBank database (36).

DATABASE AVAILABILITY

The TMPad can be accessed at <http://bio-cluster.iis.sinica.edu.tw/TMPad>. All of the fully annotated entries in the TMPad can be downloaded in the XML (extensible markup language) format. The database content is updated monthly and manually reviewed. The web pages of the TMPad were implemented in PHP and the jQuery JavaScript Library (37) on Apache web server. All data in the TMPad are managed by MySQL database system (38). The images of static 3D structures were generated using PyMol (39) while the visualization of 3D structures was achieved by using the Jmol package (40).

DATABASE ACCESS

User interfaces

The TMPad enables extensive searches and provides user-friendly interfaces for retrieving different levels of helix-packing related information about a TM protein. We show the summary of a typical workflow in the TMPad with an example in Figure 1. In this example we show the steps in accessing the information by searching for a PDB identifier or the corresponding keywords. The ‘Overview’ of the retrieved entry is first displayed. A quick overview of the protein including the lists of the complete and individual TM domain chains, experimental details and cross references is shown. The structural and functional annotations from GO and SCOP databases of the protein chains are also shown. The ‘Topology’ information is shown in a cartoon for the selected TM chain and we show the sequence, classification by HELANAL, orientation and computed tilt angles of each TM helix with respect to membrane. In the ‘Helix–helix Interactions’ page, users are first presented with a summary of all helical pair interactions of the selected chain, along with a helix–helix interaction graph in a top view where the nodes and the edges represent the TM helices and the pairwise interactions, respectively. Each interaction can be further selected and expanded for retrieving the detailed geometric descriptors of helix-packing described in the previous section on ‘Database Content’, in addition to the visualization in Jmol. Furthermore, in the ‘Lipid accessibility’ page, the helix–helix interaction graph is combined with RLA information represented by helical wheels. The edges of interaction graph are labeled with the number of contacts and crossing angles between the interacting helical pair. Each helical wheel shows a color-coded gradient scale of the calculated RLA value for each residue ranging from 0 (dark red: very buried) to 1 (dark blue: very exposed). In addition, the global surface of the protein is displayed in wireframes and both interior (highlighted in yellow) and exterior (in

TMPad TRANSMEMBRANE PROTEIN HELIX-PACKING DATABASE

Home | Search | PDB ID: **1xio** | Go

Keyword Search
sensory rhodopsin | All fields
AND | Anabaena | All fields

Overview

Protein 1XIO

PDB identifier	1XIO				
Title	Anabaena sensory rhodopsin				
Chain	PDB	PDBTM	UniProt AC	GO	
	A	A	Q8YSC4 (Q8YSC4_ANASP)	Show GO	
SCOP	Chain A: Bacteriorhodopsin-like...				

Topology

1XIO Chain A

Chain A contains 7 helices

Label	N-	Start	Sequence	End	C-	Length	Tilt	Class
H1	out	7	LHWIYVAGMTIGALHFWSLRN	28	in	22	19.20°	Curved
H2	in	35	YEYLVAMFPIWWSGLAYMAMA	55	out	21	3.32°	Curved
H3	out	70	YARYIDWMVTTPLLLLSSLWTA	91	in	22	6.78°	Kink
H4	in	99	WTLLGFLMSTQIVVITSGLIAD	120	out	22	11.57°	Curved
H5	out	128	RYLWYICGVCAFLILWGIWNPLR	151	in	24	12.59°	Linear
H6	in	166	DKLVTYFTVLWIGYPIVWIIG	186	out	21	16.87°	Curved
H7	out	198	DTFLFCLLPFFSKVGFSFLDLHG	220	in	23	16.08°	Kink

Helix-helix Interactions

1XIO Chain A

Chain A contains 7 helix-helix interactions

Helix 1	Helix 2	Type	Ω angle	D _{min}	CT	HII
A:H1	A:H2	A/LH	-163.03°	4.75	17	Jmol

A:H1 - A:H2

Anti-parallel/Left-handed

Crossing angle	-163.03°
Rotational angle α	-95.78°
Rotational angle β	34.33°
A:H1 scalar shift	2.93 Å
A:H2 scalar shift	-5.12 Å

Minimal distance (Å)	
Ca distance	4.75
Atom distance	2.62
Axis distance	8.30

Lipid Accessibility

Chain A Helix 1 Lipid Accessibility

Ligands

RET	1	Ligand/Ion	RETINAL
Name	Seq.Id	Binding Site	
RET	301 (A)	TRP A: 76, THR A: 80, LEU A: 83, VAL A: 112, GLY A: 116, TYR A: 132, PRO A: 180, TRP A: 183, LYS A: 210	

Figure 1. The summary of the user interfaces and a typical workflow of analysis in the TMPad with an example by searching for an existing PDB identifier or specific keywords. Entering ‘1xio’ or ‘Anabaena sensory rhodopsin’ in the search directs the result to the ‘Overview’ page where the summary of experimental details, functional and structural classifications, cross references to other databases is provided. For ‘Topology’ information, the TM helices are defined in terms of sequence, orientation, tilt and geometric characterization. ‘Helix-helix Interactions’ can be examined first in a summary table and expanded to display detailed helix-packing features. The visualization can be displayed centered on each interacting helical pair. The ‘Lipid Accessibility’ page supplements helix-helix interactions with calculated lipid accessibility information with helical wheels and interaction graph, as well as global visualization of surface and cavities. Lastly, the ‘Ligands’ page offers visualization of binding sites to ligands. Note that users can freely browse through any of the categories for a selected protein in any order.

green) cavities can also be viewed in Jmol. Lastly, in the ‘Ligands’ view we list the ligands bound to all protein chains in a table with the contact residues. For visualization of ligands, we enabled a customized menu for viewing the selected ligand in Jmol. Users can select different options for visualizing ligands and their contact side-chains, or adjust parameters for viewing the protein–ligand or ligand–ligand complexes.

Search options and an example of application

There are several search options for accessing the information in the TMPad. As mentioned above and shown in Figure 1, users can enter a valid PDB identifier or keywords to begin the search. However, users are not limited to this option alone. To further illustrate the database functionalities, we discuss the following example as an application to study helix-packing in TM proteins. In this example, we assume the user is investigating and comparing the topologies, helix-packing interactions and ligand-binding sites of a specific protein family such as the aquaporins (AQPs). To retrieve a list of ‘AQP-like’ structures in the database, we limited the keyword search to the term ‘aquaporin’ and selected representatives from different types of archaeal, bacterial and eukaryotic origins. When there were multiple structures found for a particular type, we selected the one structure solved by X-ray with the best resolution. The following representative structures were selected for this analysis: bovine AQP0 (PDB ID: 1ymg_A) (41), bovine AQP1 (PDB ID: 1j4n_A) (42), human AQP4 (PDB ID: 3gd8_A) (43), human AQP5 (PDB ID: 3d9s_A) (44), archaeal AQPM (PDB ID: 2f2b_A) (45), *Escherichia coli* AQPZ (PDB ID: 2o9g_A) (46), yeast AQY1 (PDB ID: 2w2e_A) (47), spinach PIP2 (plasma membrane intrinsic protein; PDB ID: 3cn5_A) (48) and *E. coli* GlpF (aquaglyceroporin; PDB ID: 1ldf_A) (49). We downloaded the data for all nine proteins and compared them based on different levels of information in the TMPad.

Several interesting observations can be made. First, although the above ‘AQP-like’ proteins are diverse in origin, substrates, or tissues expressed, they share striking similarities in their topologies and helix-packing. With respect to the membrane topology, the examined AQP-family proteins share the same overall architecture, with six TM helices and two re-entrant loops arranged in a pseudo 2-fold symmetry. The reentrant loops have been well-characterized for its important role in substrate recognition during translocation (50). This process is largely facilitated by the essential Asn-Pro-Ala (NPA) motif and its adjacent Arg residue located in the two re-entrant loops. We show the NPA motifs in AQPM and GlpF protein complexes with glycerol molecules in Figure 2a and b, respectively. It can be observed that the glycerol molecule is in contact with Asn199 (Figure 2a) or Asn203 (Figure 2b) of the NPA motifs in the re-entrant regions buried in the core of the channels. Several hydrophobic amino acids on other TM helices also participate in these interactions.

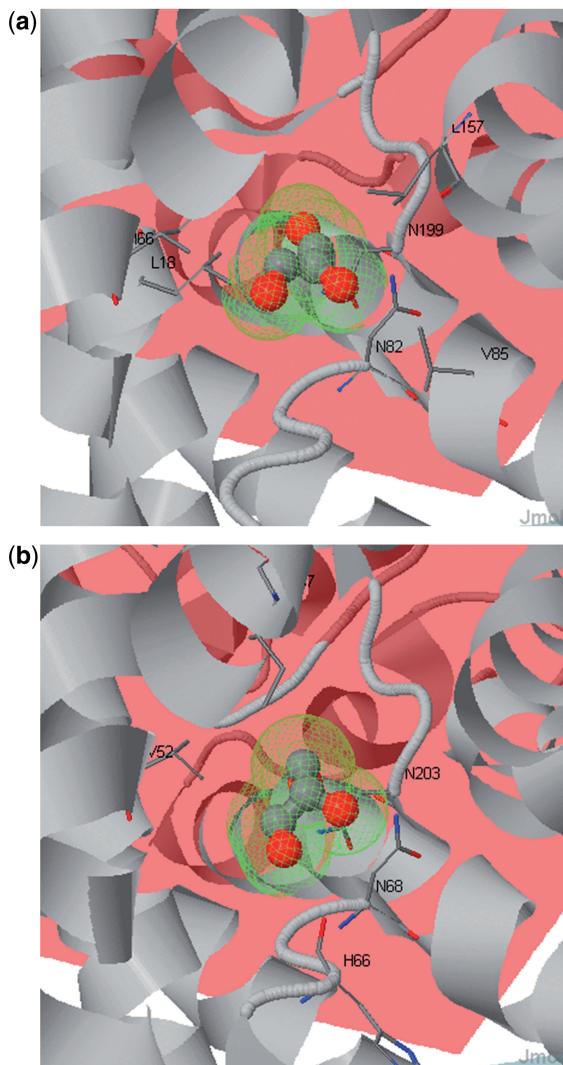


Figure 2. (a) The ligand-binding site of AQPM (PDB ID: 2f2b_A) structure complex with glycerol. The NPA motif which is important in substrate recognition can be seen with Asn199 located on the reentrant loop and several hydrophobic residues forming a ‘hydrophobic patch’. (b) The ligand-binding site of GlpF (PDB ID: 1ldf_A) structure complex with glycerol. Shown for comparison with Figure 2a, Asn203 occupies the NPA motif in the reentrant loop.

In addition, the examined proteins exhibit similar global helix–helix interactions patterns, resembling a ‘horseshoe-shaped’ network topology. We show the helix–helix interaction graphs for all proteins in Supplementary Figure S2. For brevity, only the helix–helix interaction topology of AQP0 supported by at least three VDW contacts is shown in Figure 3a. Each monomer of the above proteins has a total of five identical helix–helix interactions (TM1/TM2, TM1/TM3, TM2/TM5, TM4/TM5 and TM4/TM6). We further compared the proteins on the basis of their constituent helix–helix interactions. It can be observed that the same interacting helical pairs across all proteins show similar preferences toward their handedness (right-handed), crossing angles (with minimum and maximum standard deviations [SD_{\min} , SD_{\max}] = [2.08°,

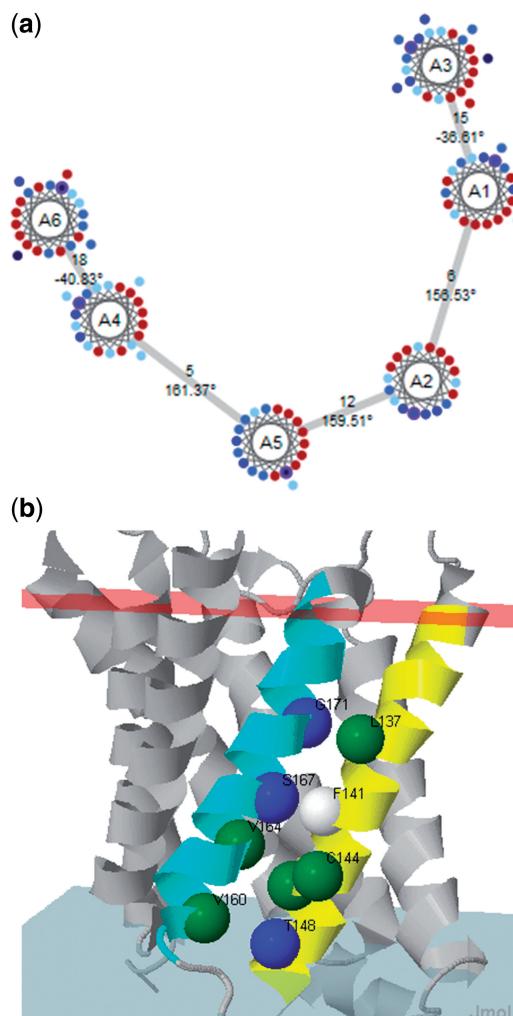


Figure 3. (a) The ‘horseshoe-like’ topology of helix–helix interaction patterns supported by at least three VDW contacts and their lipid accessibility in AQPs. For brevity, only the helix–helix interaction graph of AQP0 (PDB ID: 1ymg_A) is shown. The nodes and the edges represent TM helices and the interactions, respectively. The number of contacts and the crossing angle between each pair are also labeled and each residue in the helical wheel is colored in a gradient from red (very buried) to blue (very exposed) based on the RLA values. (b) The structure and highlighted interacting helical pair TM4 (yellow) and TM5 (light blue) within AQP0 (PDB ID: 1ymg_A) with labeled hydrophobic (green), polar (blue) and aromatic (white) contacts. The red and grey planes indicate the extracellular and cytoplasmic membrane boundaries, respectively.

4.42°]) and minimum distances between the backbone C_α atoms ($SD_{\min} = 0.15\text{\AA}$; $SD_{\max} = 0.45\text{\AA}$). The details of all helix–helix interaction comparison are listed in Supplementary Table S1. Interestingly, we also observed that TM4/TM5 utilized a higher number of hydrophobic (not including Ala) contacts than the others which were mainly composed of small (including Ala and Gly) and polar amino acids. For example, we show the hydrophobic contacts between TM4/TM5 of AQP0 in Figure 3b. Using a multiple sequence alignment program, M-coffee (51), we aligned the sequences of all proteins and observed that the hydrophobic residues (I, L, V) were concentrated on TM5

located one turn away from one of the conserved residues, His, whose key role was previously characterized for the selectivity filter in AQPs (50). However, AQPM and GlpF were exceptions to the above observation, where Gly and Ile were found at this position, respectively. The output of the multiple sequence alignment is in Supplementary Figure S3. It can be observed that some residues of the ‘hydrophobic patch’ are localized on TM5 in all proteins and they tend to be buried rather than exposed to the lipids. Our observation is in agreement with recent characterization of the hydrophobic residues exerting an additional effect on selectivity filter by lowering the permeability of hydrophobic solutes (52).

TMPad statistics

As of 10 October 2010, the TMPad contains 896 α -helical proteins, 2685 chains of TM domains and 10 289 helix–helix interactions. The vast majority of the proteins were solved by X-ray diffraction (85.5%) and solution NMR (11.4%). As for the origin of proteins, the top two groups include bacterial and eukaryotic species by taxonomy, accounting for 50 and 31% of the records in the TMPad, respectively. Furthermore, we also show the distribution of GO terms of the proteins containing TM chains in the TMPad according to the Cellular Component (CC), Molecular Function (MF) and Biological Process (BP) terms as annotated by GO in Figure 4. We calculated these distributions based on Level 2 terms that are immediate child nodes of the root node in CC, MF and BP and we also removed redundant counts for each protein. The top terms in CC include ‘cell part’ (>80%), ‘organelle’ and ‘organelle part’ (~55%), and ‘macromolecular complex’ (36%). For MP terms, most proteins carry the labels in ‘transporter activity’ (55%), followed by ‘binding’ (46%), ‘catalytic activity’ (39%) and ‘electron carrier activity’ (30%). Lastly, terms such as ‘cellular process’ (67%), ‘localization’ (64%) and ‘metabolic process’ (57%) are among the most overrepresented terms in BP. Furthermore, we also calculated and compared the amino acid compositions of the whole protein or TM domains in the TMPad as shown in Figure 5. As expected, the TM domains contain higher fractions in hydrophobic residues (maximum difference in Leu of 5.5%) and a sharp decrease for charged residues (difference of 2.6–3.4%). There is also a slightly lowered composition of polar residues and an increase in overall Phe composition (2.9%) in the TM domains, but no significant difference for Trp and Tyr.

With respect to helix-packing statistics, we show the distribution of helix–helix crossing angles (Ω) in Figure 6 and that of helical tilt angles (τ) in Figure 7. The majority (42%) of interacting helical pairs in TMPad is packed in left-handed and anti-parallel configuration ($-180^\circ < \Omega < -110^\circ$). This group also has a relatively narrower spread ($SD = 13^\circ$) with a mean and a median between 155° and 160° . The remaining groups in descending order of size are: right-handed and anti-parallel ($110^\circ < \Omega < 180^\circ$; 25%), right-handed and parallel ($-70^\circ < \Omega < 0^\circ$; 18%) and lastly left-handed and

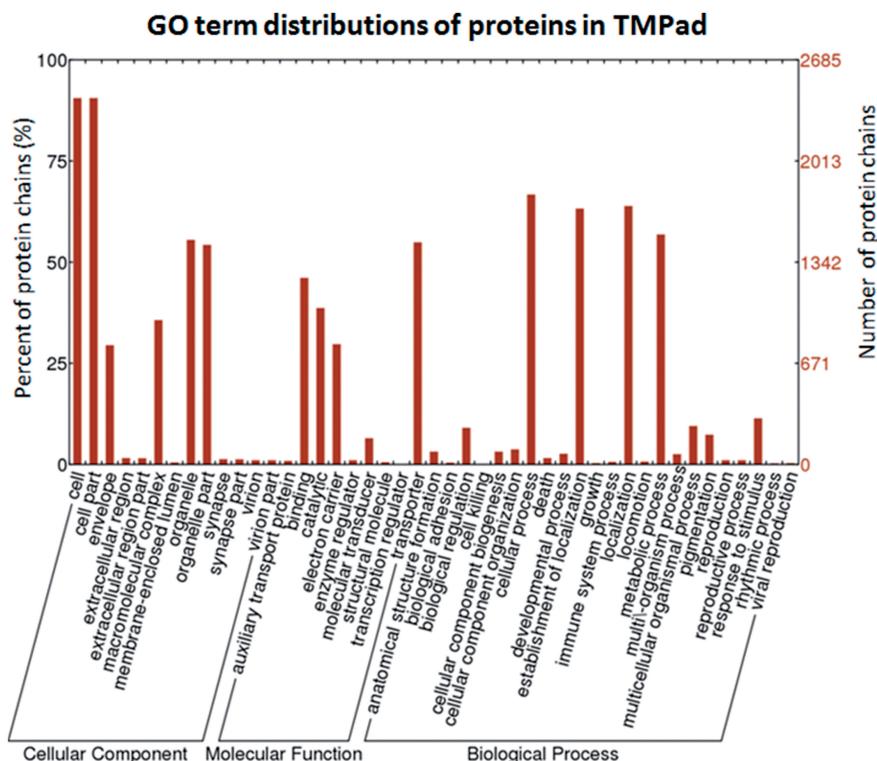


Figure 4. The plot of GO term distributions of protein chains in TMPad based on Level 2 terms in the ‘CC’, ‘MF’ and ‘BP’ categories. Each term is labeled on the X-axis and the percent and number of protein chains are labeled on the left and right side of the Y-axis, respectively. The plot was drawn using the WEGO (Web Gene Ontology Annotation Plot) tool (56).

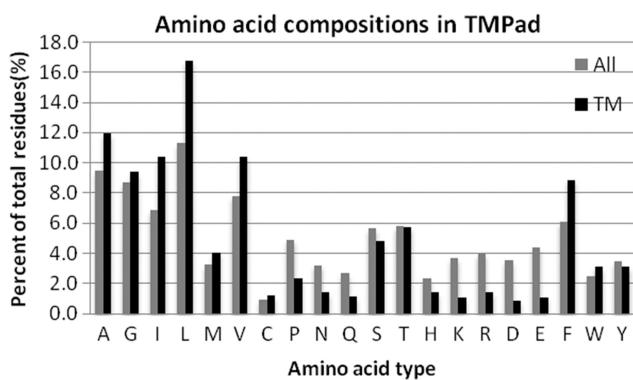


Figure 5. The comparison of amino acid compositions of all proteins based on the whole protein (blue bars) and exclusively TM domains (red bars) in the TMPad. There is a higher composition of hydrophobic residues, especially Leu (5.5%), in the TM domains compared to overall composition, and otherwise for most polar and charged residues. In addition, Phe shows a higher fraction (2.8%) in TM domains compared to other aromatic residues.

parallel helical pairs ($0^\circ < \Omega < 70^\circ$; 13%). In Figure 7, the distribution of tilt angles of all helices in TMPad appears slightly left-skewed, with a median of 22° , a mean of 23° and a SD of 11° . Lastly, the largest the helical geometry class is the ‘curved’ type (53%), while the ‘kinked’ type represents ~37% of all helices. Only ~9% of all helices are classified as ‘linear’ based on the current entries of the database.

CONCLUSION

The TMPad database is a comprehensive source for studying helix-packing and it is aimed to provide leverage for examining among different levels of structural information in topology, lipid accessibility, and ligand-binding information in α -helical TM proteins. As we have shown in the example, the thorough helix-packing descriptors we provide in the TMPad may be investigated on their own terms and/or in conjunction with the above structural features for better understanding of their relationships. In addition, there are other applications of the database particularly in structural modeling such as the derivation of knowledge-based potentials, and structure prediction. For the former problem, it was shown that crossing angles and contact pairs can be used to formulate new energy functions (53). In addition, two recent works have used interacting helical pairs as template library for reconstructing helical bundles *in silico* (54,55). Thus, the TMPad may be used for not only gaining insights into helix-packing in existing structures, but also has broad application in development of new structure prediction methods. Lastly, with the recent advances in structure determination in membrane proteins, we anticipate the rate of solved structures to gain momentum. To further enhance the TMPad, we plan in the future to develop a web server for calculating helix-packing based on user input. This work will be a continued and joint effort with structural biologists and computer scientists for

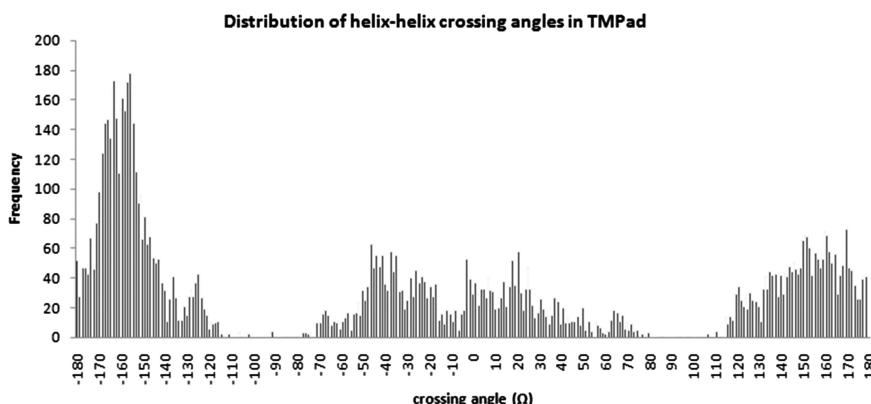


Figure 6. The distribution of crossing angles of interacting helical pairs in a histogram. On the *X*-axis is the observed crossing angle of each helix-helix interaction by bins, and the *Y*-axis represents the frequency in each bin.

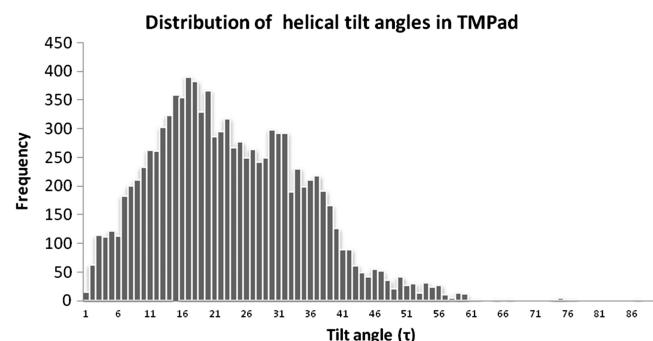


Figure 7. The distribution of helical tilt angles in a histogram. On the *X*-axis is the observed tilt angle of each helix by bins, and the *Y*-axis represents the frequency in each bin.

expanding the database functionality and information content.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Chao-Ling Lin for designing the logo of the website.

FUNDING

The thematic program of Academia Sinica (AS95ASIA03); National Science Council, partial (NSC97-2627-P-001-004). Funding for open access charge: Academia Sinica Investigator Award, Academia Sinica.

Conflict of interest statement. None declared.

REFERENCES

- Vinothkumar,K.R. and Henderson,R. (2010) Structures of membrane proteins. *Q. Rev. Biophys.*, **43**, 65–158.
- Wallin,E. and von Heijne,G. (1998) Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.*, **7**, 1029–1038.
- Stevens,T.J. and Arkin,I.T. (2000) Do more complex organisms have a greater proportion of membrane proteins in their genomes? *Proteins*, **39**, 417–420.
- White,S.H. (2009) Biophysical dissection of membrane proteins. *Nature*, **459**, 344–346.
- Popot,J.L. and Engelman,D.M. (1990) Membrane protein folding and oligomerization: the two-stage model. *Biochemistry*, **29**, 4031–4037.
- Popot,J.L. and Engelman,D.M. (2000) Helical membrane protein folding, stability, and evolution. *Annu. Rev. Biochem.*, **69**, 881–922.
- Chothia,C., Levitt,M. and Richardson,D. (1981) Helix to helix packing in proteins. *J. Mol. Biol.*, **145**, 215–250.
- Langosch,D. and Heringa,J. (1998) Interaction of transmembrane helices by a knobs-into-holes packing characteristic of soluble coiled coils. *Proteins*, **31**, 150–159.
- Gimpelev,M., Forrest,L.R., Murray,D. and Honig,B. (2004) Helical packing patterns in membrane and soluble proteins. *Biophys. J.*, **87**, 4075–4086.
- Kim,S., Jeon,T.J., Oberai,A., Yang,D., Schmidt,J.J. and Bowie,J.U. (2005) Transmembrane glycine zippers: physiological and pathological roles in membrane proteins. *Proc. Natl Acad. Sci. USA*, **102**, 14278–14283.
- Adamian,L. and Liang,J. (2002) Interhelical hydrogen bonds and spatial motifs in membrane proteins: polar clamps and serine zippers. *Proteins*, **47**, 209–218.
- Walters,R.F. and DeGrado,W.F. (2006) Helix-packing motifs in membrane proteins. *Proc. Natl Acad. Sci. USA*, **103**, 13658–13663.
- Beuming,T. and Weinstein,H. (2004) A knowledge-based scale for the analysis and prediction of buried and exposed faces of transmembrane domain proteins. *Bioinformatics*, **20**, 1822–1835.
- Yin,H., Slusky,J.S., Berger,B.W., Walters,R.S., Vilaire,G., Litvinov,R.I., Lear,J.D., Caputo,G.A., Bennett,J.S. and DeGrado,W.F. (2007) Computational design of peptides that target transmembrane helices. *Science*, **315**, 1817–1822.
- Tusnady,G.E., Dosztanyi,Z. and Simon,I. (2005) PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res.*, **33**, D275–D278.
- Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
- Tusnady,G.E., Dosztanyi,Z. and Simon,I. (2005) TMDet: web server for detecting transmembrane regions of proteins by using their 3D coordinates. *Bioinformatics*, **21**, 1276–1277.
- Lomize,M.A., Lomize,A.L., Pogozheva,I.D. and Mosberg,H.I. (2006) OPM: orientations of proteins in membranes database. *Bioinformatics*, **22**, 623–625.

19. Tusnady,G.E., Kalmar,L. and Simon,I. (2008) TOPDB: topology data bank of transmembrane proteins. *Nucleic Acids Res.*, **36**, D234–D239.
20. Burba,A.E., Lehnert,U., Yu,E.Z. and Gerstein,M. (2006) Helix Interaction Tool (HIT): a web-based tool for analysis of helix-helix interactions in proteins. *Bioinformatics*, **22**, 2735–2738.
21. Rose,A., Goede,A. and Hildebrand,P.W. (2010) MPPlot—a server to analyze and visualize tertiary structure contacts and geometrical features of helical membrane proteins. *Nucleic Acids Res.*, **38**(Suppl.), W602–W608.
22. Marsico,A., Scheubert,K., Tuukkanen,A., Henschel,A., Winter,C., Winnenburg,R. and Schroeder,M. (2010) MeMotif: a database of linear motifs in alpha-helical transmembrane proteins. *Nucleic Acids Res.*, **38**, D181–D189.
23. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
24. Andreeva,A., Howorth,D., Chandonia,J.M., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.
25. Laskowski,R.A. (2009) PDBsum new things. *Nucleic Acids Res.*, **37**, D355–D359.
26. Reichert,J. and Suhnel,J. (2002) The IMB Jena Image Library of Biological Macromolecules: 2002 update. *Nucleic Acids Res.*, **30**, 253–254.
27. The UniProt Consortium. (2009) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **37**, D169–D174.
28. Bansal,M., Kumar,S. and Velavan,R. (2000) HELANAL: a program to characterize helix geometry in proteins. *J. Biomol. Struct. Dyn.*, **17**, 811–819.
29. Lo,A., Chiu,H.S., Sung,T.Y., Lyu,P.C. and Hsu,W.L. (2008) Enhanced membrane protein topology prediction using a hierarchical classification method and a new scoring function. *J. Proteome Res.*, **7**, 487–496.
30. Bernsel,A., Viklund,H., Hennerdal,A. and Elofsson,A. (2009) TOPCONC: consensus prediction of membrane protein topology. *Nucleic Acids Res.*, **37**, W465–W468.
31. Jones,D.T. (2007) Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics*, **23**, 538–544.
32. Lee,J. and Im,W. (2007) Implementation and application of helix-helix distance and crossing angle restraint potentials. *J. Comput. Chem.*, **28**, 669–680.
33. Pappu,R.V., Marshall,G.R. and Ponder,J.W. (1999) A potential smoothing algorithm accurately predicts transmembrane helix packing. *Nat. Struct. Biol.*, **6**, 50–55.
34. Lee,B. and Richards,F.M. (1971) The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.*, **55**, 379–400.
35. Hubbard,S.J. and Thornton,J.M. (1993) ‘NACCESS’ Computer Program. Department of Biochemistry and Molecular Biology, University of College London.
36. Wishart,D.S., Knox,C., Guo,A.C., Cheng,D., Shrivastava,S., Tzur,D., Gautam,B. and Hassanali,M. (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, **36**, D901–906.
37. The jQuery JavaScript Library. <http://jquery.com> (29 November 2010, date last accessed).
38. The MySQL database. <http://www.mysql.com> (29 November 2010, date last accessed).
39. DeLano,W.L. (2002) The PyMOL Molecular Graphics System on World Wide Web. <http://www.pymol.org> (29 November 2010, date last accessed).
40. Jmol. an open-source Java viewer for chemical structures in 3D. <http://www.jmol.org> (29 November 2010, date last accessed).
41. Harries,W.E., Akhavan,D., Miercke,L.J., Khademi,S. and Stroud,R.M. (2004) The channel architecture of aquaporin 0 at a 2.2-A resolution. *Proc. Natl Acad. Sci. USA*, **101**, 14045–14050.
42. Sui,H., Han,B.G., Lee,J.K., Walian,P. and Jap,B.K. (2001) Structural basis of water-specific transport through the AQP1 water channel. *Nature*, **414**, 872–878.
43. Ho,J.D., Yeh,R., Sandstrom,A., Chorny,I., Harries,W.E., Robbins,R.A., Miercke,L.J. and Stroud,R.M. (2009) Crystal structure of human aquaporin 4 at 1.8 Å and its mechanism of conductance. *Proc. Natl Acad. Sci. USA*, **106**, 7437–7442.
44. Horsefield,R., Norden,K., Fellert,M., Backmark,A., Tornroth-Horsefield,S., Terwisscha van Scheltinga,A.C., Kvassman,J., Kjellbom,P., Johanson,U. and Neutze,R. (2008) High-resolution x-ray structure of human aquaporin 5. *Proc. Natl Acad. Sci. USA*, **105**, 13327–13332.
45. Lee,J.K., Kozono,D., Remis,J., Kitagawa,Y., Agre,P. and Stroud,R.M. (2005) Structural basis for conductance by the archaeal aquaporin AqpM at 1.68 Å. *Proc. Natl Acad. Sci. USA*, **102**, 18932–18937.
46. Savage,D.F. and Stroud,R.M. (2007) Structural basis of aquaporin inhibition by mercury. *J. Mol. Biol.*, **368**, 607–617.
47. Fischer,G., Kosinska-Eriksson,U., Aponte-Santamaría,C., Palmgren,M., Geijer,C., Hedfalk,K., Hohmann,S., de Groot,B.L., Neutze,R. and Lindkvist-Petersson,K. (2009) Crystal structure of a yeast aquaporin at 1.15 angstrom reveals a novel gating mechanism. *PLoS Biol.*, **7**, e1000130.
48. Nyblom,M., Frick,A., Wang,Y., Ekwall,M., Hallgren,K., Hedfalk,K., Neutze,R., Tajkhorshid,E. and Tornroth-Horsefield,S. (2009) Structural and functional analysis of SoPIP2;1 mutants adds insight into plant aquaporin gating. *J. Mol. Biol.*, **387**, 653–668.
49. Tajkhorshid,E., Nollert,P., Jensen,M.O., Miercke,L.J., O’Connell,J., Stroud,R.M. and Schulten,K. (2002) Control of the selectivity of the aquaporin water channel family by global orientational tuning. *Science*, **296**, 525–530.
50. Gonen,T. and Walz,T. (2006) The structure of aquaporins. *Q. Rev. Biophys.*, **39**, 361–396.
51. Moretti,S., Armougom,F., Wallace,I.M., Higgins,D.G., Jongeneel,C.V. and Notredame,C. (2007) The M-Coffee web server: a meta-method for computing multiple sequence alignments by combining alternative alignment methods. *Nucleic Acids Res.*, **35**, W645–W648.
52. Hub,J.S. and de Groot,B.L. (2008) Mechanism of selectivity in aquaporins and aquaglyceroporins. *Proc. Natl Acad. Sci. USA*, **105**, 1198–1203.
53. Im,W., Lee,J., Kim,T. and Rui,H. (2009) Novel free energy calculations to explore mechanisms and energetics of membrane protein structure and function. *J. Comput. Chem.*, **30**, 1622–1633.
54. Barth,P., Wallner,B. and Baker,D. (2009) Prediction of membrane protein structures with complex topologies using limited constraints. *Proc. Natl Acad. Sci. USA*, **106**, 1409–1414.
55. Harrington,S.E. and Ben-Tal,N. (2009) Structural determinants of transmembrane helical proteins. *Structure*, **17**, 1092–1103.
56. Ye,J., Fang,L., Zheng,H., Zhang,Y., Chen,J., Zhang,Z., Wang,J., Li,S., Li,R. and Bolund,L. (2006) WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res.*, **34**, W293–W297.